CLAIMS

1. A computerized method for automatically detecting deviations in a data table comprising a multitude of records and a multitude of columns, said method comprising the steps of:

(1) selecting a column as a classification column;

(2) executing a classification method for calculating a classification tree with respect to said classification column, whereby each edge of said classification tree is associated with a predicate, whereby a leaf node of said classification tree is associated with a leaf record set comprising a subset of records for which a class predicate comprising all predicates along a path from a root node of said classification tree to said leaf node evaluates to TRUE, and whereby said leaf node is associated with a leaf label representing an expected value in said classification column of said leaf record set; and

(3) determining from said leaf record set all records deviating with respect to said classification column from said leaf label as a deviation set.

2. The method of claim 1 wherein said deviation set is associated with said class predicate in said determining step as an explanation for being detected as a deviation.

3. The method of claim 1 wherein a multitude of leaf nodes are calculated in said executing step and wherein said determining step is executed for said multitude of leaf nodes.

4. The method of claim 3 wherein for each leaf node a purity value is determined, said purity value measuring the degree of conformity of an associated leaf record set with respect to a leaf label of said leaf node.

5. The method of claim 4 wherein said purity value is based on the percentage of the number of records of said leaf record set not coinciding with said leaf label.

6. The method of claim 4 wherein leaf nodes and their associated leaf record set with a purity value indicating a conformity below a predefined first purity threshold are disregarded.

7. The method of claim 6 wherein leaf nodes and their associated leaf record set with purity values indicating a conformity above a predefined second purity threshold are also disregarded.

8. The method of claim 7 wherein said purity thresholds have values between 80% and 100%.

9. The method of claim 1 wherein in said executing step said classification tree is limited to a depth not exceeding a predefined depth threshold.

10. The method of claim 9 wherein said depth threshold is a number not exceeding 3.

11. The method of claim 1 wherein said method is iterated by repeatedly executing said method for another column of said data table as a classification column.

12. The method of claim 1 wherein for each leaf node and its corresponding deviation set a ranking value is determined and wherein said deviation sets are ordered in a sequence according to their ranking values.

13. The method of claim 12 wherein said ranking value is determined based on the purity in value of a leaf node.

14. The method of claim 12 wherein said ranking value is determined based on the number of records associated with a leaf node.

15. The method of claim 12 wherein said ranking value is determined based on the length of the class predicate of a leaf node.

16. The method of claim 12, wherein for each pair of first and second deviation sets with a non-empty intersection set, said intersection set is treated as a third deviation set with a ranking value higher than a ranking value of said first and second deviation sets and said third deviation set is associated with a class predicate comprising the combination of the class predicates of said first and second deviation sets.

17. The method of claim 1 wherein, if said data table has been modified after said method has been executed in a first run, the following further steps are performed:

executing said method in a second run with respect to the modified data table;

determining said deviation set of said modified data table; and

reducing said deviation set by said deviation set of said first run.

18. A system for automatically detecting deviations in a data table comprising a multitude of records and a multitude of columns, said system comprising means for carrying out the steps of the method of claim 1.

19. A data processing program for execution in a data processing system comprising software code portions for performing the method of claim 1 when said program is run on a computer.

20. A computer program product stored on a computer usable medium, comprising computer readable program means for causing a computer to perform a the method of claim 1 when said program is run on said computer.